Inferences for two samples
Math 218, Mathematical Statistics
D Joyce, Spring 2016

General questions for statistical inferences. We've seen some of the most commonly used point estimators, interval estimates, and confidence intervals. There are many more, and we'll look at some of them. There are a few general questions to keep in mind when you see a new method for a statistical inference.

- Under what circumstances does the method apply? What kind of sample or samples are under consideration? What are the hypotheses that the method assumes.
- If it's a hypothesis test, what are the null and alternative hypotheses?
- What is the test statistic? How is it computed from the sample? What kind of distribution does it have?
- How do you apply the method?

Two designs are used when there are two sample populations. Chapter 8 is about comparing means of two populations, one usually being a treatment group, the other a control group.

There are two kinds of these comparisons. One kind is the *independent samples design*, and for that the individuals in the treatment group are different than the individuals in the control group. In fact, the sizes of the two populations don't even have to be the same. The other kind is the *matched pairs design*. Here the individuals are the same in the two populations, but the responses are different, for instance, the control group data may be measured before an experiment, and the treatment data measured after the experiment, but on the same n subjects.

In the first, the independent samples design, we simply have two separate populations and we sample each independently. We assume that the population distributions are of the same kind but have different unknown parameters. The sample sizes, n_1 and n_2, for the two populations don't have to be the same but they may be, and it makes sense to have them the same size unless there's some good reason that one should be smaller.

The second, the matched pairs design, also has two populations, but the samples are not independent. We'll look at that case next time. The sample sizes for the two populations are the same, and the i^{th} random variable X_i from the first population is not assumed to be independent from i^{th} random variable Y_i from the second population, although they're assumed to be independent of all the other random variables. For example, X_i might be a pretest score and Y_i the posttest score for the same individual.

Comparing means for two independent samples. There are various tests that apply depending on what is assumed. In all cases we assume that we have two populations where we take a sample

$$X_1, X_2, \ldots, X_{n_1}$$

of size n_1 from the first population and one

$$Y_1, Y_2, \ldots, Y_{n_2}$$

of size n_2 from the second population. (In practice, n_1 is usually equal to n_2, but it's not necessary.) Furthermore, we assume all $n_1 + n_2$ random variables are independent. We'll denote the population means as μ_1 and μ_2, and the population variances as σ_1^2 and σ_2^2. Our job is to make inferences about the population means and/or variances.

Comparing means for large samples. Here we assume that the sample sizes n_1 and n_2 are large. In this case the sample means $\overline{X}$ and $\overline{Y}$ are approximately normally distributed with means μ_1 and μ_2 and variances σ_1^2/n_1 and σ_2^2/n_2, respectively. Therefore, the difference of the means, $\overline{X} - \overline{Y}$, is also normally distributed and its mean is the difference of the means, $\mu_1 - \mu_2$, while the variance of is the sum of the variances, $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Therefore,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has an approximately standard normal distribution.

Since we're assuming that the sample sizes are large, we can replace the unknown population variances σ_1^2 and σ_2^2 by the corresponding samples variances S_1^2 and S_2^2. Therefore, our test statistic

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

has an approximately standard normal distribution.

Based on this test statistic, an α-level confidence interval for $\mu_1 - \mu_2$ has ends

$$(\overline{x} - \overline{y}) \pm z_{\alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

Therefore, a two-sided hypothesis test for $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_0 : \mu_1 - \mu_2 \neq \delta_0$, where δ_0 is a specified difference of the means, rejects H_0 if $|z| > z_{\alpha/2}$ where

$$z = \frac{(\overline{x} - \overline{y}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

Usually the queston is: do the two populations have the same mean? In that case δ_0 is 0, so you reject $H_0 : \mu_1 = \mu_2$ if $|z| > z_{\alpha/2}$ where

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

Corresponding one-sided hypothesis tests are summarized in table 8.2 of the text.

Comparing means for small samples from normal populations. If you assume the two populations are normally distributed, then the difference $\overline{X} - \overline{Y}$ is normally distributed even when the sizes of the samples are small. We'll look at one of tests designed for small samples from normal populations.

In the first test, we assume that the two populations have the same unknown variance σ^2. In that case, the test statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S\sqrt{1/n_1 + 1/n_2}}$$

has a t-distribution with $(n_1 - 1) + (n_2 - 1) = n - 2$ degrees of freedom, where S^2 is the pooled sample variance given by

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

Confidence intervals and hypotheses tests are the same for this test as above, except the test statistic has a t-distribution instead of a z-distribution.

There's a different test when the two populations are not assumed to have the same variance.

Inferences for two samples for matched pair design. Paired t-tests.

With a matched pairs design, there are two populations, but the samples are not independent. The sample sizes for the two populations are the same n, and the i^{th} random variable X_i from the first population is not assumed to be independent from i^{th} random variable Y_i from the second population, in fact, we'll assume that there is a specific correlation ρ between them. $\rho = \text{Corr}(X_i, Y_i)$, the same correlation for all i. We do assume, however, that for *different* subscripts all the random variables are independent.

We'll also assume that that the populations are normally distributed, $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$. Then, as in the independent samples design, their differences

$$D_i = X_i - Y_i$$

are independent and normally distributed with mean $\mu_D = \mu_1 - \mu_2$ and variance

$$\begin{aligned}\sigma_D^2 = \mathrm{Var}(D_i) &= \mathrm{Var}(X_i - Y_i) \\ &= \mathrm{Var}(X_i) + \mathrm{Var}(Y_i) - 2\mathrm{Cov}(X_i, Y_i) \\ &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\end{aligned}$$

Since $D_1, \ldots, D_n$ is sample from a normal distribution $N(\mu_1 - \mu_2, \sigma_D^2)$ where $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$, we can replace σ_D^2 by the sample variance S_D^2 to get a T-statistic and then use the t-intervals and tests we studied in the last chapter. It follows, for instance, that two-sided $100(1-\alpha)\%$ confidence level interval for μ_D has endpoints

$$\overline{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

where $\overline{d}$ is the sample average $\frac{1}{n}\sum d_i$ which equal $\overline{x} - \overline{y}$, and s_d^2 is the sample variance $\frac{1}{n-1}\sum(d_i - \overline{d})^2$.

As an example hypothesis test, to test the null hypothesis $H_0 : \mu_D = \delta_0$ (where typically δ_0 is 0), against the alternative hypothesis $H_1 : \mu_D \neq \delta_0$, reject H_0 when the t-statistic

$$t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{n}}$$

has an absolute value greater than $t_{n-1,\alpha/2}$, which is equivalent to the condition

$$|\overline{d} - \delta_0| > t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}.$$

Hypothesis tests like this are called *paired t-tests*.

The advantage of this matched pairs test over the corresponding independent samples test is that the sample variance of the differences, s_d^2, will be much smaller than the pooled sample variance, $s_1^2/n_1 + s_2^2/n_2$. This will make the interval estimates shorter and the the hypothesis tests finer.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`